Language Analysis of American Government Documents

By: Griffin Talan

Data Science Problem

- *What is fake news?*
- *Why is the news so biased?*
- *What is the correct response to Covid?*

- How has the way we speak changed over time and how has the found its way into political speech? If there is a trend, can this be seen across agencies within our government?
    - Has overall sentiment and level of subjectivity changed?

Agenda

- Outline Data Sourcing and Preparation
- Discuss data processing
- Discussing biderectional LSTM modeling
- Discuss model predictions for prediction datasets and relevant trends
- Discuss overall trends
- Discuss limitations and next steps involved
- Demo Heroku App / HTML prediction app

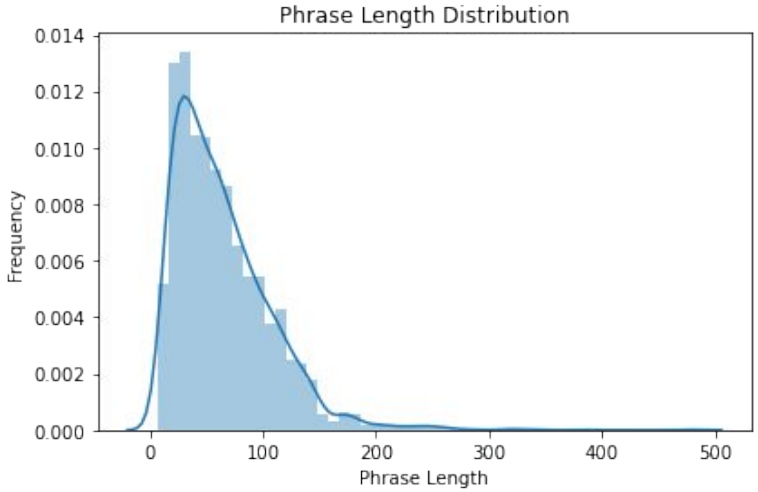
Training Data

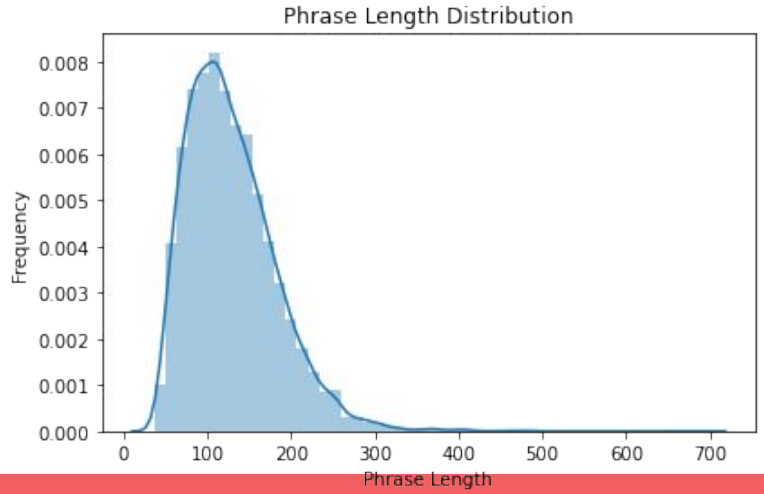
Sentiment

- UCI Machine Learning dataset of combined IMDB, Amazon, and Yelp reviews
- 500 positive and 500 negative reviews for each data set

Subjectivity

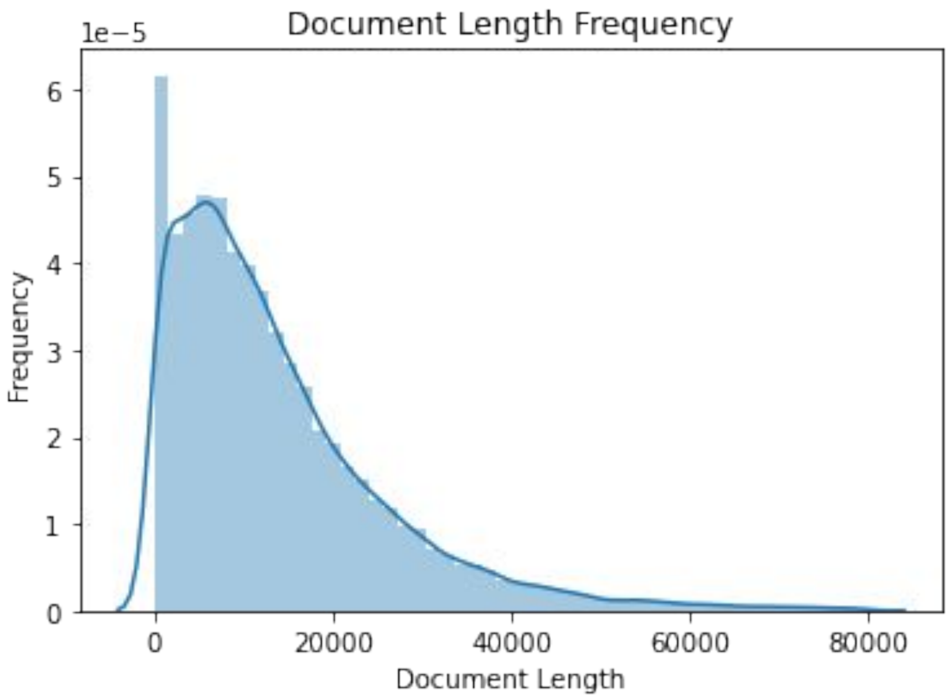
- Cornell University annotated rotten tomatoes movies review dataset
- 5000 subjective and 5000 objective labelled sentences





Prediction Data

- State of the Union Addresses
  - Scraped using BeautifulSoup
  - 240 addresses
- Inaugural Addresses
  - 58 addresses
- DOJ Press Releases
  - ~13,000 records
- Supreme Court Decisions
  - ~35,000 decisions



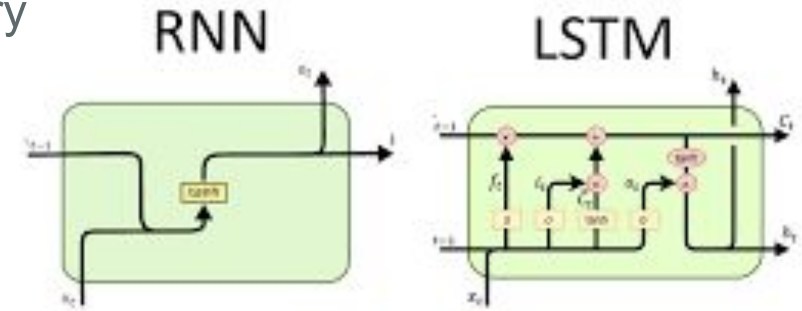
Methods - Text Preprocessing

- Lowercased
- Removed non-alpha characters
- Tokenized by words
- Option to remove stop words and stem data - neither were used
- Truncation to a maximum number of words and padding extra space
- Converted text to numeric data using tensorflow tokenizers fit with word vocab from training data
  - Stored as pickle file and hosted in app with model

Methods - Bidirectional LSTM Model

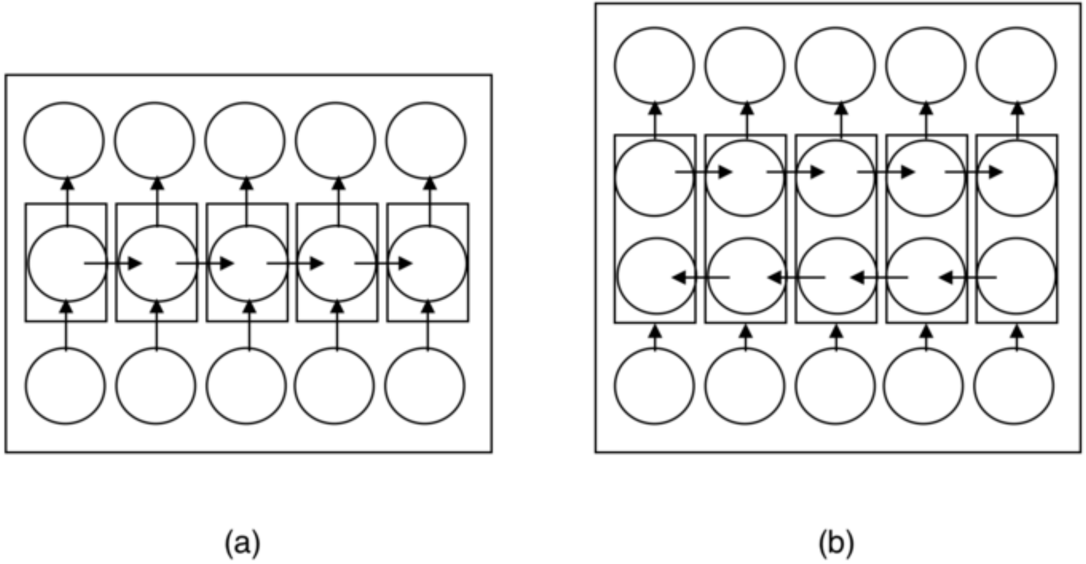
LSTM uses 3 “gates” to regulate cell memory

- Input
- Forget
- Output



Methods - Bidirectional LSTM Model

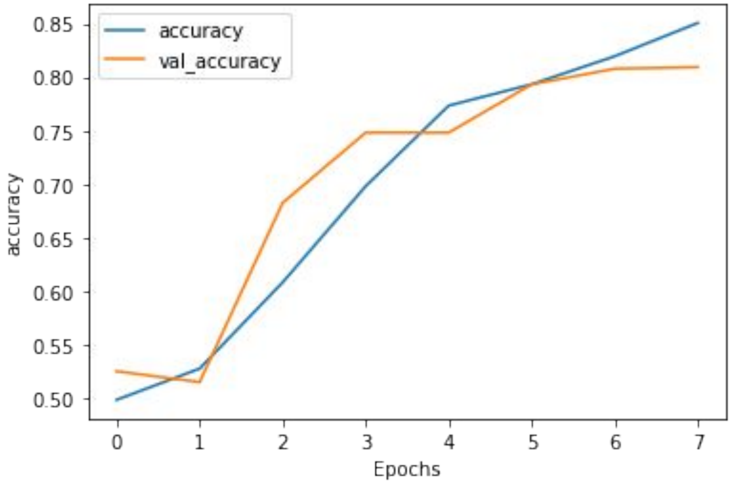
Bidirectional LSTM models use input from future and past cells to address weighting

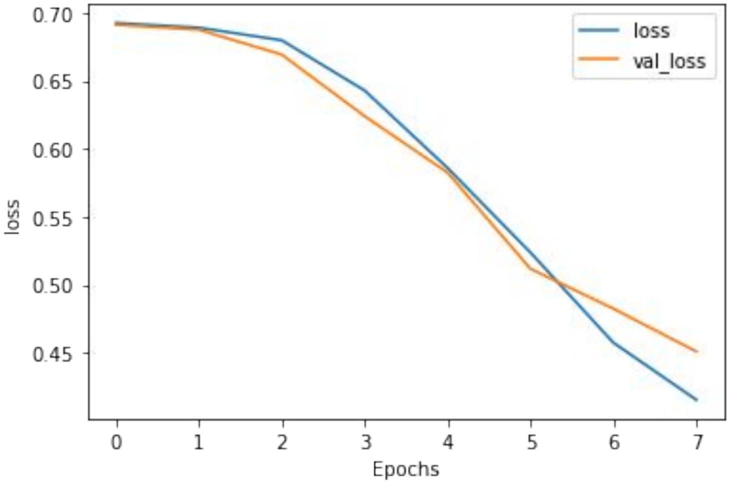


Structure overview
(a) unidirectional RNN
(b) bidirectional RNN

Model Training - Sentiment Analysis

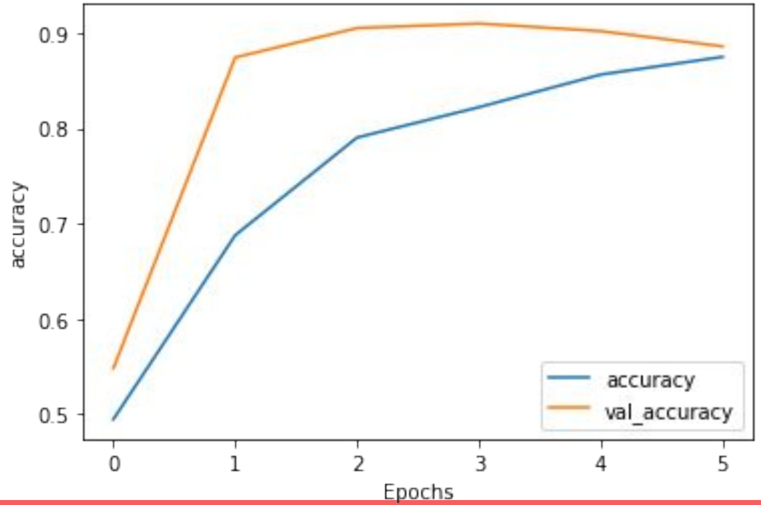
- Training was highly regularized to address overfitting
  - Reduced NN layer nodes
  - Added high dropout after every step (>0.7)
- Accuracy ~0.80 for train and test

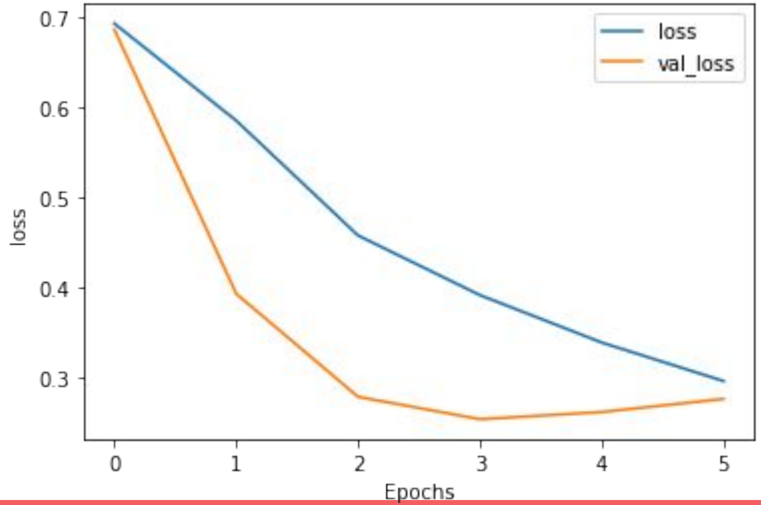




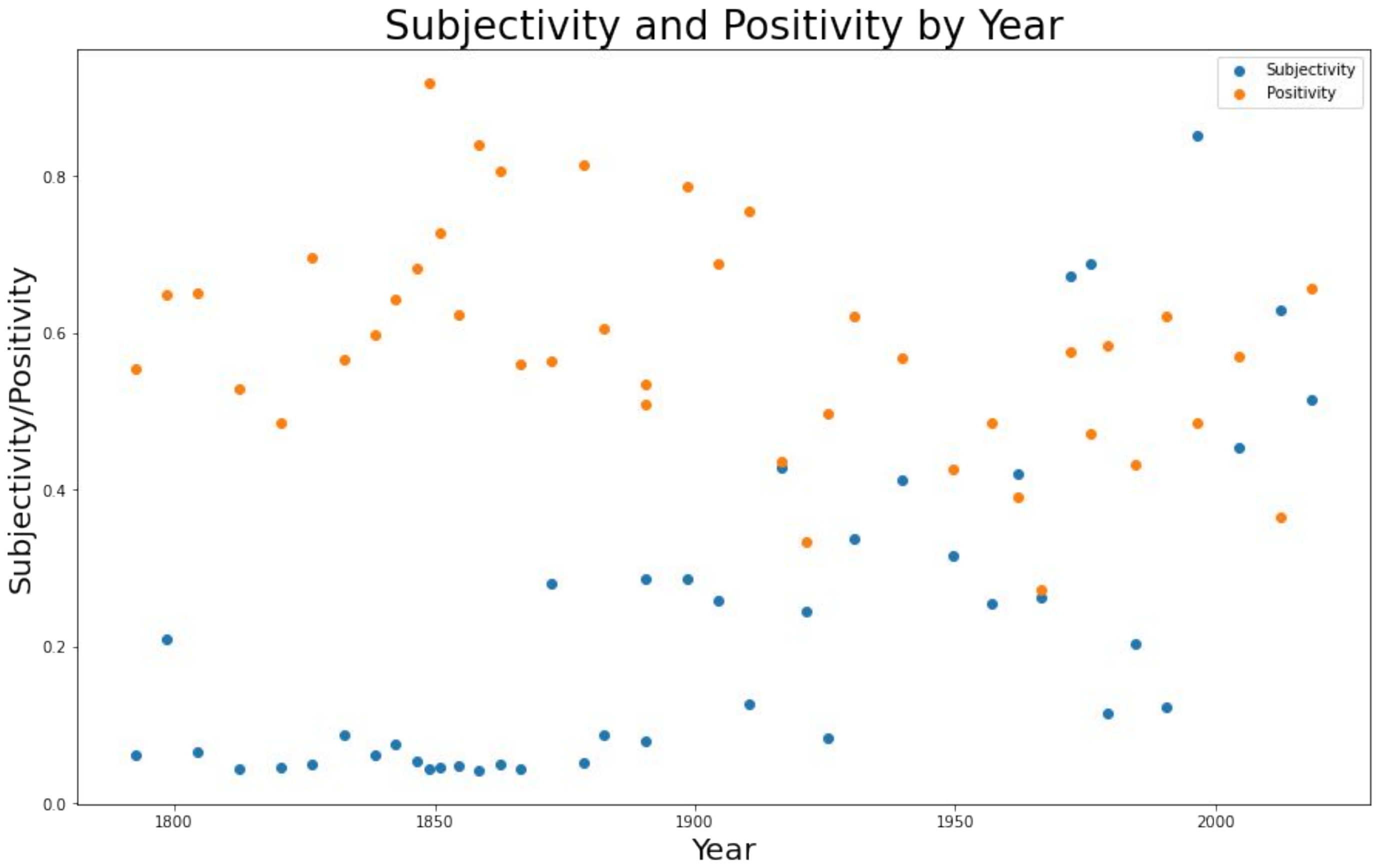
Model Training - Subjectivity Analysis

- Training was highly regularized to address overfitting
  - Reduced NN layer nodes
  - Added high dropout after every step (>0.7)
- Accuracy ~0.85 for train and test



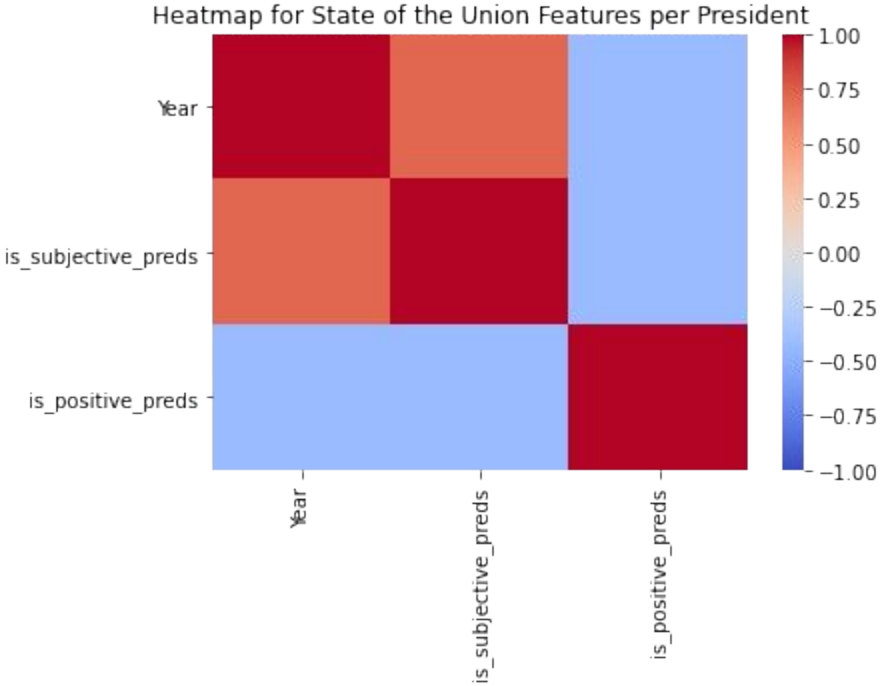


Predictions - State of the Union Addresses



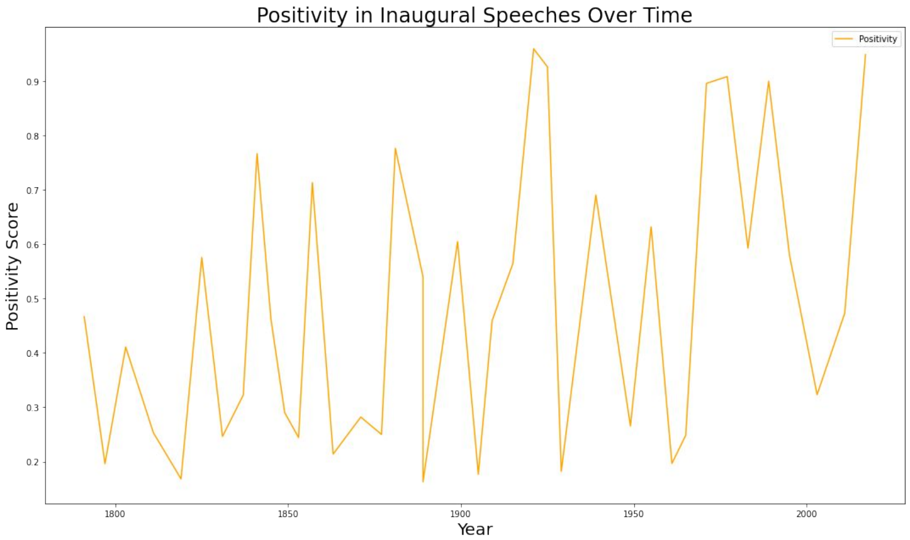
Predictions - State of the Union Addresses

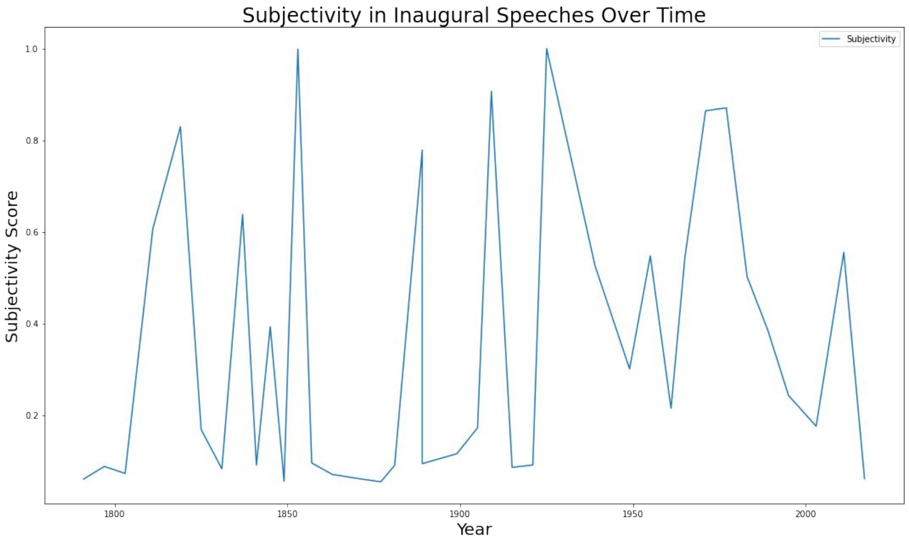
- Positive correlation between year and subjectivity rating
- Minor negative correlation between positivity



Predictions - Inaugural Addresses

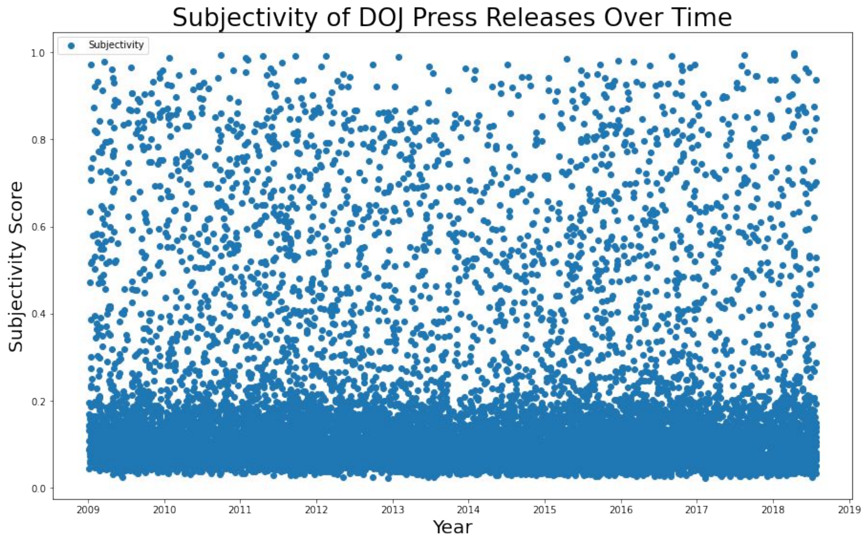
- No visible trends were found in the inaugural addresses
  - Fewest number of documents, infrequently occurs

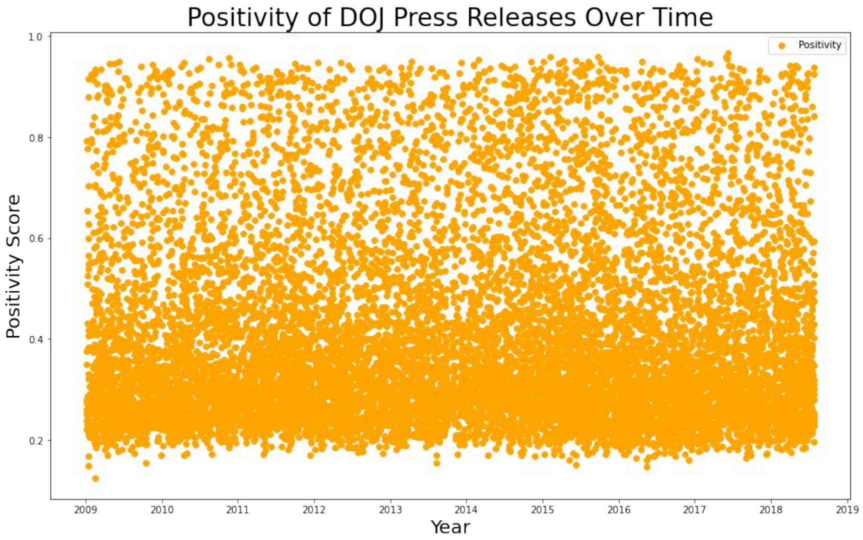




Department of Justice

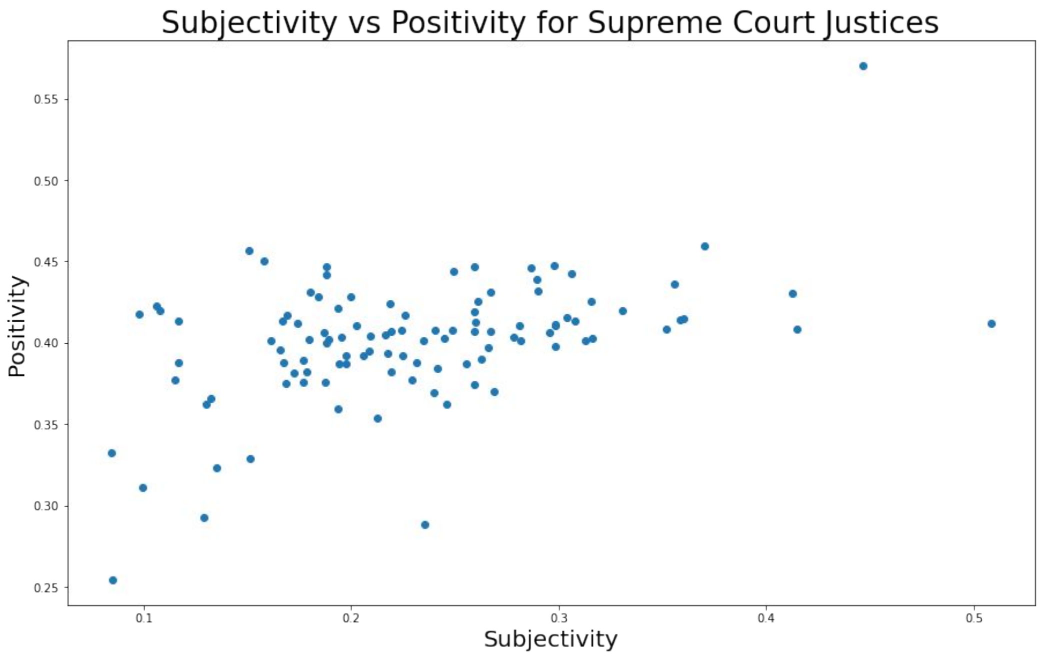
- The DOJ data only covers the last 10 years
- Shows consistent, objective, non-positive tone





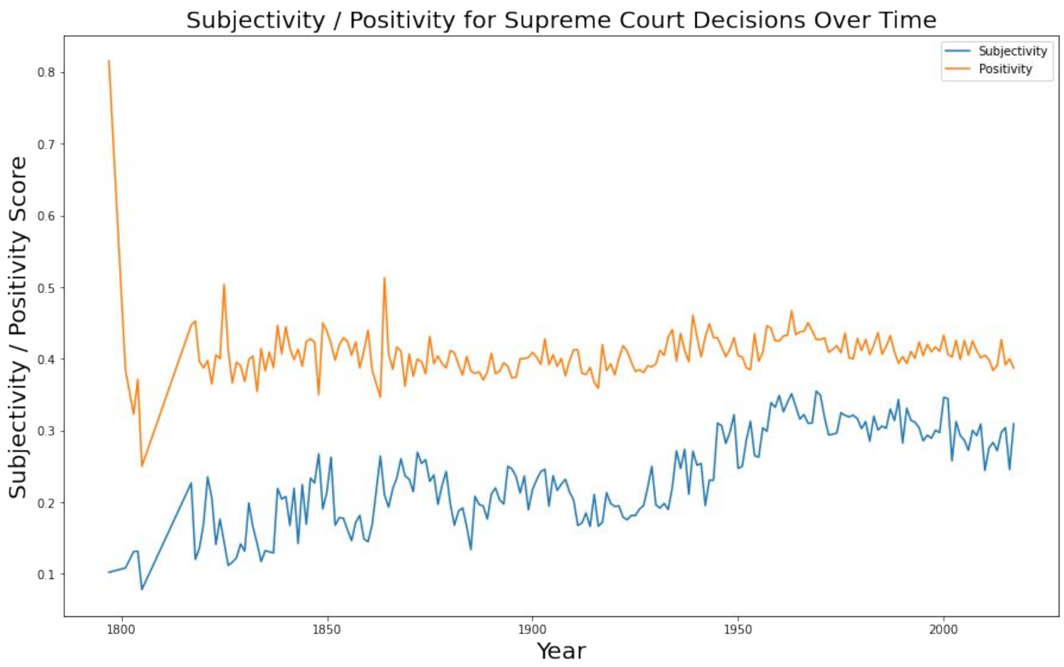
Supreme Court Decisions

- After removing decisions by judges with only one decision, trends were noted in documents written per judge



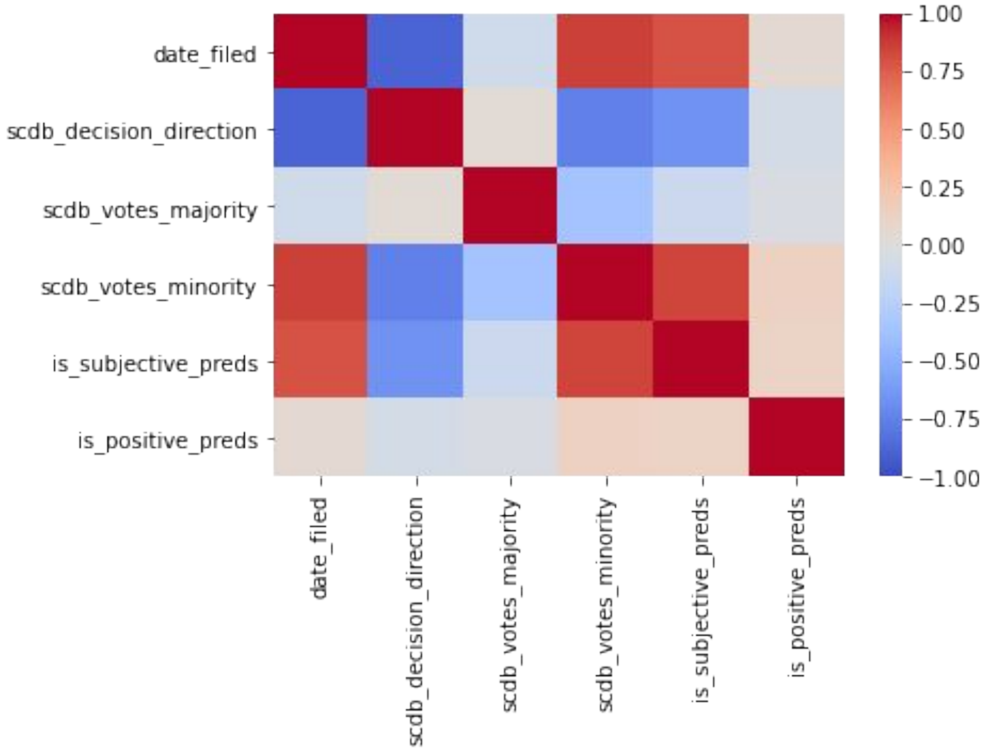
Supreme Court Decisions

- Generally stable sentiment state
- Increase in subjectivity over time



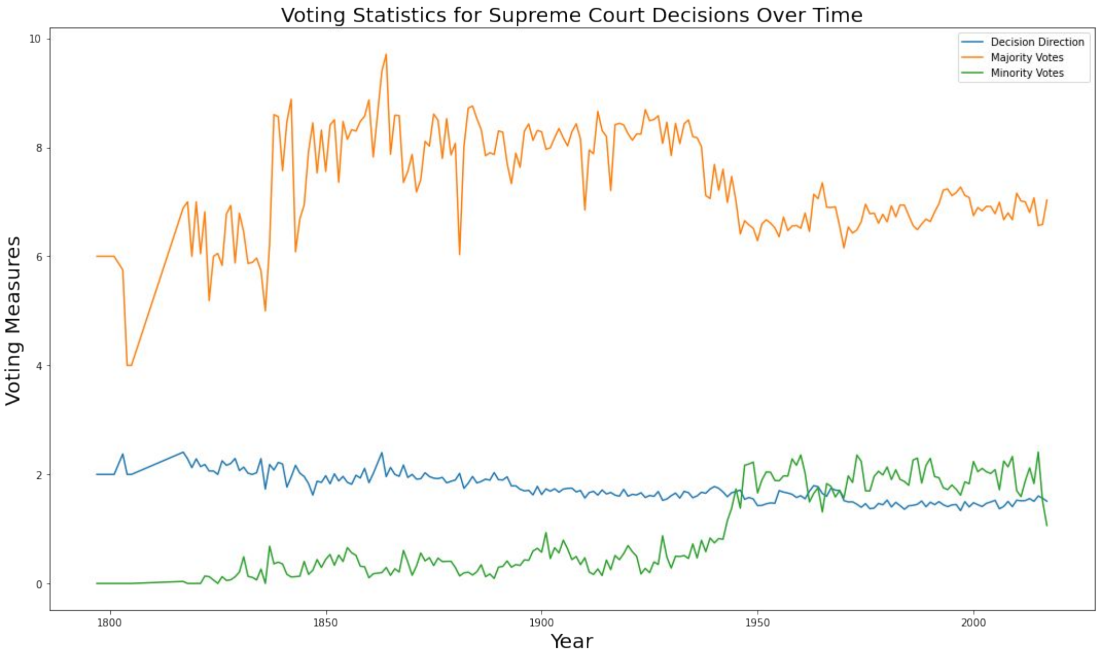
Supreme Court Decisions

- Correlation between date filed and subjectivity
- Correlation between date filed and minority vote count



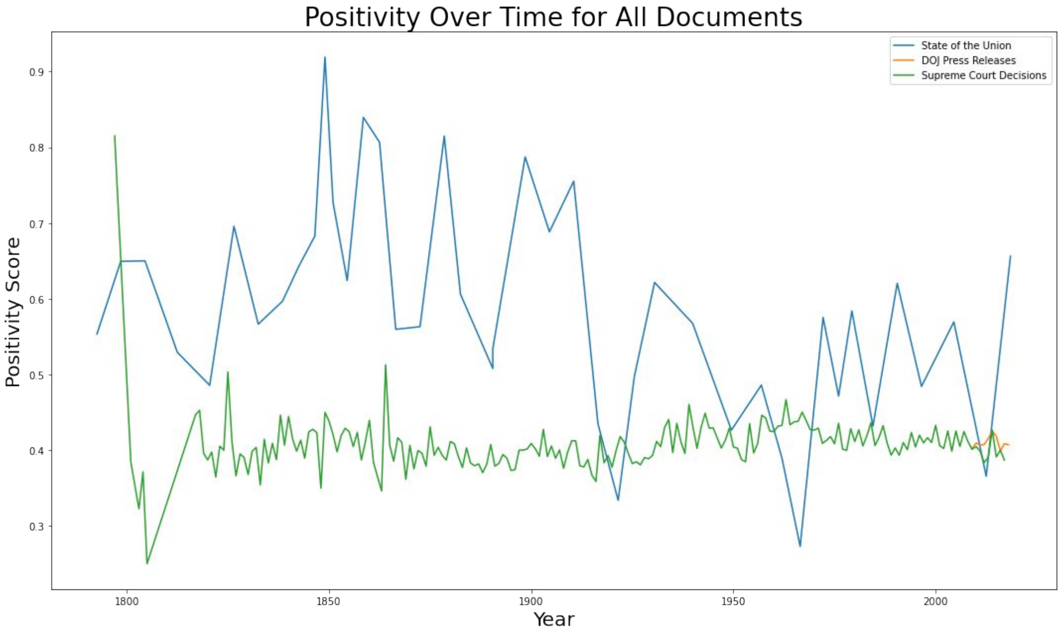
Supreme Court Decisions

- Shift from majority-dominated courts to more evenly split after World Wars



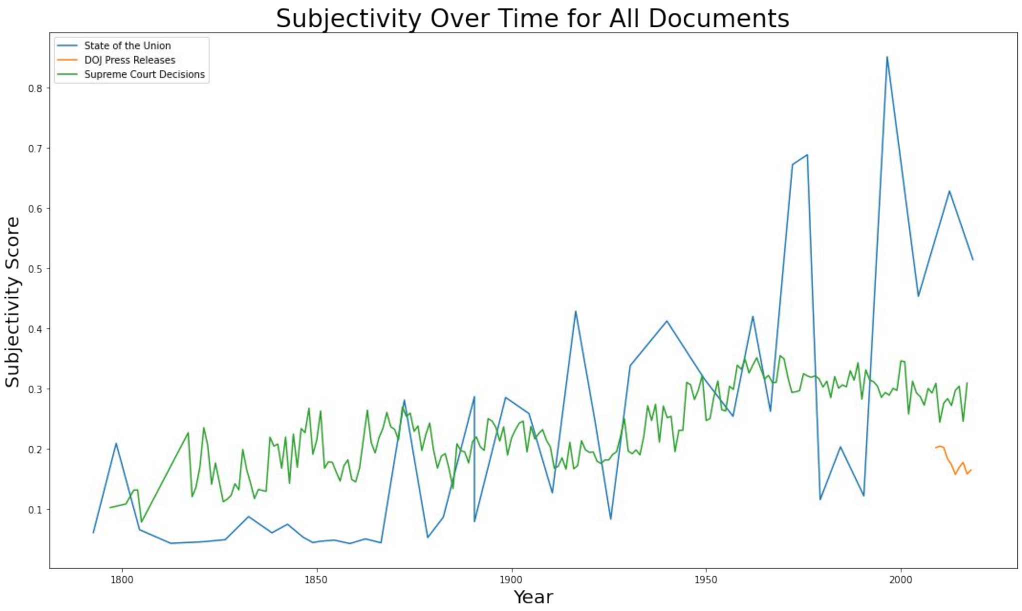
Sentiment Overlay

- Mostly stable across all documents analyzed
- Some negative correlation seen in the State of the Union Speeches



Subjectivity Overlay

- Increase in subjectivity correlated with time in both State of the Union and Supreme Court decisions



Concerns and Next Steps

- Get better training data that spans larger ranges of times
  - Ensemble method that can include year in prediction model
- Redo the predictions on a by-sentence basis for the documents and average
- Different training data - very fit to purpose
  - Train data is customer reviews, not speeches toward constituencies
  - Does not have wide enough vocabulary to account for everything in political documents
- Topic labeling / keyword extraction

Heroku Demo

https://griffin-subj-sent.herokuapp.com

References

1. Sentiment Data
    a. 'From Group to Individual Labels using Deep Features', Kotzias et. al,. KDD 2015
2. Subjectivity Data
    a. Bo Pang and Lillian Lee, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, *Proceedings of ACL 2004*.
3. State of the Union Speeches
    a. https://www.presidency.ucsb.edu/documents/presidential-documents-archive-guidebook/annual-messages-congress-the-state-the-union#Table%20of%20SOTU
4. Inaugural Addresses
    a. http://www.bartleby.com/124/
5. DOJ Press Releases
    a. https://www.kaggle.com/jbencina/department-of-justice-20092018-press-releases
6. Supreme Court Decision
    a. http://scdb.wustl.edu/data.php
7. LSTM Cell Image
    a. http://dprogrammer.org/rnn-lstm-gru
8. Bidirectional LSTM Structure
    a. https://en.wikipedia.org/wiki/Bidirectional_recurrent_neural_networks